DOCUMENT RESUME

ED 409 320 TM 026 545

AUTHOR Wang, Lin; Fan, Xitao
TITLE The Effect of Cluster Sampling Design in Survey Research on the Standard Error Statistic.
PUB DATE Mar 97
NOTE 17p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Cluster Analysis; Educational Research; *Error of Measurement; Estimation (Mathematics); *Research Design; *Sampling; Simulation; *Surveys
IDENTIFIERS Jackknifing Technique; *Variance (Statistical)

ABSTRACT

Standard statistical methods are used to analyze data that is assumed to be collected using a simple random sampling scheme. These methods, however, tend to underestimate variance when the data is collected with a cluster design, which is often found in educational survey research. The purposes of this paper are to demonstrate how a cluster design affects the standard error statistic and the subsequent analyses, and to present practical techniques to analyze data from cluster designs correctly. A heuristic example is given to illustrate how to compute the variance estimate for a cluster design and the corresponding design effect. Simulation data is then used to examine variance estimation results from one- and two-stage cluster designs, respectively. Both a formula approach and a jackknife resampling approach are used in obtaining variance estimates. It is shown that, for 150 observations sampled from a population of 1,000, using a 2-stage cluster design, the actual variance can be underestimated by a factor of 3 if the standard statistical method is used. The underestimated variance or standard error statistic will lead to unwarranted statistical significance in hypothesis testing, or a narrow confidence interval in parameter estimation. Consequently, misleading conclusions can be made based on these inappropriate analysis findings. (Contains 1 tables and 11 references.) (Author/SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED 409 320

PERMISSION TO REPRODUCE AND
DISSEMINATE THIS MATERIAL
HAS BEEN GRANTED BY
Lin Wang

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER-(ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)
This document has been reproduced as
received from the person or organization
originating it.
☐ Minor changes have been made to
improve reproduction quality.
• Points of view or opinions stated in this
document do not necessarily represent
official OERI position or policy.

# THE EFFECT OF CLUSTER SAMPLING DESIGN IN SURVEY RESEARCH ON THE STANDARD ERROR STATISTIC

Lin Wang
ACT

Xitao Fan
Utah State University

Paper presented at the Annual Meeting of the American Educational Research Association, March, 1997, Chicago.



BEST COPY AVAILABLE

TM026545

## Abstract

Standard statistical methods are used to analyze data that is assumed to be collected using a simple random sampling scheme. These methods, however, tend to underestimate variance when data is collected with a cluster design, which is often found in educational survey research. The purpose of this paper is to (a) demonstrate how a cluster design affects standard error statistic and the subsequent analyses, and (b) present practical techniques to correctly analyze data from cluster designs. A huristic example is given to illustrate how to compute the variance estimate for a cluster design and the corresponding design effect. Simulation data is then used to examine variance estimation results from one- and two-stage cluster designs, respectively. Both formula approach and a jackknife resampling approach are used in obtaining variance estimates. It is shown that, for 150 observations sampled from a population of 1,000, using a two-stage cluster design, the actual variance can be underestimated by a factor of 3 if the standard statistical method is used. The underestimated variance or standard error statistic will lead to unwarranted statistical significance in hypothesis testing, or a narrow confidence interval in parameter estimation. Consequently, misleading conclusions can be made based on these inappropriate analysis findings.

# The Effect of Cluster Sampling Design in Survey Research on the Standard Error Statistic

Cluster sampling design is frequently implemented in educational survey researh. Most of today's national public databases in education, such as the National Assessment of Educational Progress (NAEP), contain information collected with some types of cluster sampling designs. Cluster sampling designs are also employed by invididual researchers who conduct survey studies at smaller scales. For example, in an evaluation of the sample designs reported in an educational research journal, it was found that about 15% of the probability sampling designs involved some cluster selections (Wang, 1996). Cluster sampling is often used mainly to reduce the cost of conducting a survey. "Cluster sampling is an effective design for obtaining a specified amount of information at minimum cost under the following conditions: (1) A good frame listing population elements either is not available or is very costly to obtain, while a frame listing clusters is easily obtained. (2) The cost of obtaining observations increases as the distance separating the elements increases" (Scheaffer, Mendenhall, & Ott, 1990, p. 244). Such conditions are quite typical of educational survey research, particularly the research on organizational charactersitics of natural clusters, sucy as classes, schools, etc..

Analysis of the survey data collected with a cluster sampling design requires special treatment in variance estimation. In most statistical analyses, one important assumption for applying various statistical methods is that observations are independently and identically distributed in the population. The formulas for statistical computations are usually provided for analyzing data collected with a simple random sampling design. This

assumption, however, creates problems for analyzing data from a cluster design in survey research, because the assumption is rarely met in practical survey situation (Cochran, 1977; Deming, 1960; Jaeger, 1988; Kish, 1965; Kott, 1991; Lee, Forthofer, & Lorimor, 1989). People tend to have more in common within a natural group or cluster, such as a class, a school, a community, etc., than among the clusters, this is particularly true of opinions or attitudes. In survey analysis, this within-cluster similarity is called homogeneity and measured by rho, the coefficient of intraclass correlation (Kish, 1965).

Generally, the homogeneity tends to increase the variance of a cluster sample for a given sample size when compared with a simple random sample. The effect of such increased variance is measured by a quantity design effect, or deff. The deff is the ratio of the actual variance estimate of a cluster sample to the variance estimated from a simple random sample on the same data. Typically, the deff is greater than unity, indicating that a cluster design has less efficiency than a simple random design. In other words, given the same sample size, a cluster sample tends to yield a larger variance than does a simple random sample. This is the price we pay for using a cluster design for economy and easiness in conducting a survey study.

Because of such possible differences in the estimated variances, it is important that correct variance estimation method be used in analysis of cluster sample data. This point is not well taken by many educational researchers doing survey research. Consequently, variance estimation in the analysis of cluster sample data is usually done using the formulas appropriate for simple random sample data only (Wang, 1996). This practice entails possible underestimation of the variances, or the standard error statistics. The

consequences can be unwarranted findings some nonexistant statistical significance in hypothesis testing. For example, Kish (1965) has demonstrated that, given the same data, a cluster sample variance can be underestimated by a ratio of 4 if the formula for a simple random design is used. Very different results can be obtained from the two methods. Such a problem has hardly been discussed in educational research literature and needs to be addressed with no further delay. The purpose of this paper is to: (a) demonstrate how a cluster design affects standard error statistic and the subsequent analyses, and (b) present practical techniques to analyze data from cluster designs correctly.

## Methods and Data

The investigation of the effects of cluster sampling designs on standard error statistics is presented and discussed in two ways: an illustrative example with a small heuristic data set and a simulation study.

The heuristic data analysis. A small set of data from a cluster design is analyzed to demonstrate the differences in the results of standard error or variance calculation. This is done by using a standard statistics formula for simple random sample data and a formula appropriate for a cluster design.

The simulation study. The simulation study is designed to model the practice of taking a cluster sample and doing analyses in educational survey research. A population data set is generated such that the population contains 1,000 observations. For ease of discussion, equal cluster size is adopted for both the population and the sample in the study. Specifically, there are 20 clusters (e.g., schools) with 50 elements (e.g., teachers)

in each cluster. The data is assumed to be measurement taken on a 5-point Likert scale (say, teachers' job satisfaction rating). The clusters are defined by varying cluster means and standard deviations in data generation. This is based on the reasoning that the elements (teachers) in a cluster (school) tend to have similar characteristics (job satisfaction rating), yielding different means and standard deviations across clusters.

Both a one-stage probability cluster design and a two-stage probability cluster design are implemented. Ten clusters are randomly selected from the 20 clusters in the population data. For the one-stage design, all the data in each of the 10 clusters is used, i.e., $\underline{k} = 10$, $\underline{n} = 500$. For the two-stage design, 15 cases are randomly selected from each of the 10 cluster. Therefore, the final sample size is $\underline{n} = 150$ and $\underline{k} = 10$. The analysis and the calculation of standard errors are carried out with two approaches: formula approach and resampling approach. The formulas for variance calculation in cluster designs are readily found in sampling texts (Kish, 1965; Sheaffer et al., 1990). For example, in this study, the variance estimation formulas are the following, respectively, for the one-stage cluster sample (Kish, 1965, p. 151, (5.2.3)),

$$var(\bar{y}) = (1-\frac{a}{A})\frac{s_a^2}{a} \qquad \text{where} \qquad s_a^2 = \frac{1}{a-1}\sum_{\alpha}^{a}(\bar{y}_\alpha - \bar{y})^2$$

and for the two-stage cluster sample (Kish, 1965, p. 167, (5.6.5)),

$$var(\bar{y}) = (1-\frac{a}{A})\frac{s_a^2}{a} + (1-\frac{b}{B})\frac{a}{A}\frac{s_b^2}{ab} = \frac{s_a^2}{a} - \frac{1}{A}[s_a^2 - (1-\frac{b}{B})\frac{s_b^2}{b}]$$

where

$$s_b^2 = \frac{1}{a(b-1)} \sum_{\alpha}^{a} \sum_{\beta}^{b} (y_{\alpha\beta} - \bar{y}_{\alpha})^2$$

and

$$s_a^2 = \frac{1}{a-1} \sum_{\alpha}^{a} (\bar{y}_{\alpha} - \bar{y})^2$$

The resampling approach is used to obtain the empirical standard errors or variances instead of the theoretical ones based on the formulas given above. Resampling is a potentially very useful approach in practice where appropriate formulas are not available, such as in some comlex sample designs. A variety of resampling methods are used in statistical analysis. Bootstrap and jackknife methods appear to be very popoular. While the bootstrap method is believed to work better than the jackknife method, the latter is thought to be more suitable in cases of complex sampling (Lee et al, 1989; Mooney & Duval, 1993). In this study, we use a jackknife technique to obtain an empirical variance estimate (Efron, 1981; Efron & Gong, 1983; Lee et al, 1989). In a typical jackknife application on a sample of n observations, n rounds of calculations are done such that one observation is deleted in turn in each round, and relevant statistics (say, a mean) are obtained. The final variance of the statistic is given by the following formula (Lee et al, 1989, p. 33, (4.6)):

$$var(\bar{\bar{y}})jack=\frac{n-1}{n}\sum_{i=1}^{n}(\bar{y}_{(i)}-\bar{\bar{y}})^2 \qquad \bar{\bar{y}}=\frac{\sum_{i=1}^{n}\bar{y}_{(i)}}{n}$$

In a complex survey analysis, however, the jackknife operation is carried out on what is called primary sampling unit (PSU), which in this study is the cluster. The formula becomes

$$var(\bar{\bar{y}})jack=\frac{k-1}{k}\sum_{i=1}^{k}(\bar{y}_{(i)}-\bar{\bar{y}})^2 \qquad \bar{\bar{y}}=\frac{\sum_{i=1}^{k}\bar{y}_{(i)}}{k}$$

where k is the number of nonoverlapping subsets (clusters) in the data, each subset containing h observations (size of a sampled cluster) such that kh = n, the total observations in the entire sample. It is obvious that when h = 1, then k = n, that is, one observation is deleted at a time. In this simulation study, the jackknife method is applied to the two-stage cluster sample data, with one cluster deleted at a time. Specifically, a SAS program using macro procedures was written such that one cluster was deleted in each of the k = 10 executions. In each execution, the mean of the retained 9 clusters were calculated using PROC MEANS, a SAS statistical procedure. At the end of the 10 executions, 10 means were obtained and the jackknife variance of the means was estimated with the formula given above.

## Results and Discussions

### The heuristic example

For illustration of the difference in the variance estimation results between the two

methods for a simple random sample and a one-stage cluster sample respectively, the heuristic example in Table 1 is made to contain only 9 observations in 3 clusters selected from M clusters in a population, each population cluster size (B) is 3. That this is a one-stage cluster sampling design means that the 3 clusters are randomly sampled from M (assumed to be very large in the population) clusters and all the observations in each of the 3 clusters are included in the final sample. In other words, no sampling occurs within each cluster. Therefore, all the sampling variance comes from the between-cluster variance only. The calculation of the estimated variance for the cluster sample is shown to be .3333 in Table 1.

On the other hand, if the 9 observations are treated as from a simple random sample, the variance estimate from this simple random sample of 9 cases will be .1681. This calculation ignores the cluster structure in the data and does not take the difference between clusters into account. The corresponding design effect is then:

$$\underline{deff} = \frac{var(\bar{y})_{cluster}}{var(\bar{y})_{simple}} = \frac{.3333}{.1681} = 1.98$$

This indicates that the actual variance estimate of the cluster sample is close to twice the size that of the simple random sample, even though the data is made to differ only a little across the 3 clusters in this example. The standard errors are the square roots of the two variance estimates, .58 and .41, respectively. In a hypothesis testing situation, the standard error from the simple random sample formula is more likely than that from the cluster sample formular to lead to a statistically significant finding. By the same token, in constructing a confidence interval around the estimated parameter of interest such as a

population mean, the interval based on the standard error of .41 is narrower than the one from the standard error of .58.

---

Insert Tabel 1 about here

---

The one-stage cluster design

Using the aforementioned variance estimation formula for a one-stage cluster design, the estimated variance is .0312 for the simulation data from 500 cases in the 10 randomly selected clusters. On the other hand, when the formula for a simple random sample is used to estimate the variance, the estimate is .0016. The design effect then is: deff = .0312/.0016 = 19.5. Therefore, if one uses the standard statistical procedure to compute the variance estimate in this case, he or she will drastically underestimate the actual variance and will be very likely to reach some unsubstantiated conclusions.

The difference between the two variance estimates may be explained from two perspectives. First, from the conceptual perspective, it has already been mentioned that the difference results from whether the cluster design characteristics is kept in mind in both the design and the analysis processes. A typical problem in survey research is that researchers are often aware of the design advantage of cluster designs, but forget, ignore, or simply fail to see the special requirements in analysis. Consequently, survey data is analyzed in the same way, using the formulas for simple random samples, no matter what sampling designs are actually implemented. Second, from the analytic perspective, the difference occurs as a consequence of inappropriate decision on unit of analysis. In a cluster design, the unit of analysis is the cluster, not the individual

observations within the clusters. In a simple random design, no cluster exists and the unit of analysis is the individual observations. In this example, using n = 500 instead of k = 10 in calculation leads to the underestimated variance of .0016 instead of .0312.

The two-stage cluster design: A formula approach

In the two-stage cluster sample design, the simulation data consists of 10 clusters with 15 observations each. As was mentioned earlier, in a two-stage cluster sampling, sampling errors occur at both stages and have to be taken into consideration. Using the data for the simulated two-stage cluster design, the variance calculation formula for the two-stage design reflects this consideration as the formula below is made up of two parts:

$$var(\bar{y}) = (1 - \frac{a}{A}) \frac{s_a^2}{a} + (1 - \frac{b}{B}) \frac{s_b^2}{Ab}$$

The first part gives the between-cluster variance at stage one in selection, and the second part is for the within-cluster variance at stage two. The variance estimated with this formual is .034. The between- and within-cluster variances are .031 and .003, respectively. When the sample of 150 observations is treated as a simple random sample, however, the estimated variance is then .011. The design effect in this case is .034/.011, or deff = 3.05. This tells that using the wrong formula (the one for the simple random sample) will underestimate the actual variance by a factor of 3.

It is evident in the formula above that the number of clusters in the population, A, plays an important role in determining the final total variance. Given a fixed a, the number of clusters to select at stage one, a large A gives a small value for the selection

ratio, a/A, yielding a large between-cluster variance. The large A also minimizes the influence of the within-cluster variance in the formula. This situation occurs when a small number of clusters are randomly selected from a very large number of clusters in the population. An example would be selecting 100 schools from all the schools in the United States, or selecting 10 classes from all the classes in a large metropolitan school district.

On the other hand, the within-cluster selection ratio b/B influences the magnitude of the within-cluster variance in the formula. A large ratio means a small within-cluster variance contribution to the total variance. For example, in the results above, the within-cluster variance totals .003 only, with the within-cluster selection ratio at .3 (15/50), a rather large value from this stage of selection. Intuitively, a large selection ratio means more members of a cluster are included to yield a smaller sampling error in estimation.

The two-stage cluster design: A jackknife approach

As was described in the earlier section, the jackknife simulation was done on the k = 10 clusters. The 10 means from the 10 jackknife executions are: 3.22, 3.24, 3.24, 3.35, 3.39, 3.40, 3.34, 3.42, 3.40, 3.47. Because the cluster selection ratio, a/A is .5 (10 out of 20), the finite population correction (fpc) is also .5 (fpc = 1-a/A = 1-.5). This fpc is too large to ingore in calculation and has been included in the formula approach. Therefore, the fpc should also be used in the jackknife formula for consistency. The jackknife variance is then calculated as:

$$var(\bar{\bar{y}})_{jack} = (1-f)\left(\frac{k-1}{k}\right)\sum_{i=1}^{k}(\bar{y}_{(i)}-\bar{\bar{y}})^2$$

The variance estimate is .031, the same as the between-cluster variance estimate obtained earlier with the formula approach. It is obvious that the jackknife procedure assesses the effect of individual clusters on the results from the sample. Since the within-cluster part is not included in the jackknife procedure, only between-cluster variance estimate is used to represent the total variance estimate. This clearly underestimates the actual variance to certain extent, depending on the size of the within-cluster variance. In this example, the difference is .003, or about one tenth of the total variance estimate.

In practice, this small underestimation may not have noticeable influence on the findings. Therefore, using the jackknife procedure for this type of empirical calculation of variance may enable us to arrive at a reasonable estimate, provided the within-cluster variance is small enough to be ingored without any harm. If, in survey practice, the number of clusters in the population is very large and the number of observations taken from each selected cluster is not small, the effect of the within-cluster variance on the total variance is very likely to become inconsequential. Then the jackknife estimate should be a good approximation to the actual variance estimate.

## Conclusions

In view of the characteristics of educational survey research and the problems related to the analysis of complex survey data, the discussion in this paper may have offered some assistance in understanding the nature of the problem, i.e., the inappropriate estimation of variance as a result of ignoring the cluster sampling design. The examples used in illustration, however, are very simple ones, involving only the variances of means. This is mainly because formulas are available to calculate the

complex variance estimates if design features are known. For variance estimates of other parameters such as regression coefficients or those in multivariate analysis, there is no direct formula application for cluster variance calculation. Some epirical methods must be used to assess the effect of complex sampling on variance estimation. Not much has been reported on research in this respect and there is clearly a need for such research.

As far as most educational researchers are concerned, it is necessary to realize the problems in variance estimation raised in this paper. Essentially, when a cluster sampling design is used in data collection, the analysis should be carried out using correct variance estimation methods. This variance estimate is then used in making statistical inference, such as hypothesis testing or confidence interval construction. Failure to use the correct method is likely to lead to the underestimation of the actual variance due to the design effect. The correct methods include either the formula approach or the jackknife approach, depending on how much tracking information is available about the observations in the final sample.

References

Cochran, W. G. (1977). Sampling techniques. New York: Wiley.

Deming, W. E. (1960). Sample design in business research. New York: Wiley.

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. In Biometrika, 68(3), 589-99.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. In The American Statistician, 37(1), 36-48.

Jaeger, R. M. (1988). Survey research methods in education. In R. M. Jaeger (Ed.), Complementary methods for research in education (pp. 303-330). Washington, DC: Paper presented at the Annual Meeting of the American Educational Research Association.

Kish, L. (1965). Survey sampling. New York: Wiley.

Kott, P. S. (1991). A model-based look at linear regression with survey data. The American Statistician, 45(2), 107-112.

Lee, E. S., Forthofer, R. N., & Lorimor, R. J. (1989). Analyzing complex survey data. Newbury Park, CA: Sage.

Mooney, C.Z., & Duval, R.D. (1993). Bootstrapping: A nonparametric approach to statistical inference. Newbury Park, CA: Sage.

Scheaffer, R. L., Mendenhall, W., & Ott, L. (1990). Elementary survey sampling (4th ed.). Boston, MA: PWS-Kent.

Wang, L. (1996). A typology and evaluation of the survey sample designs in the Educational Administration Quarterly: 1980-1995. Unpublished doctoral dissertation, Texas A&M University, College Station.

Table 1

A heuristic example

| Case | Cluster (a = 3) | | | Sum |
|---|---|---|---|---|
| | 1 | 2 | 3 | |
| 1 | 1 | 2 | 3 | |
| 2 | 2 | 3 | 4 | |
| 3 | 3 | 4 | 5 | |
| Cluster total ($y_a$) | 6 | 9 | 12 | 27 |
| Cluster total squared | 36 | 81 | 144 | 261 |

Calculation:

$$var(\bar{y}) = \frac{1}{a}\frac{1}{B^2}\frac{1}{a-1}\left(\sum_{i=1}^{a} y_a^2 - \frac{y^2}{a}\right)$$

where, B = 3, the population cluster size, y is the sum of $y_a$ (Kish, 1965, p. 153, (5.2.3')). Therefore,

$$var(\bar{y}) = \left(\frac{1}{3}\right)\left(\frac{1}{3^2}\right)\left(\frac{1}{3-1}\right)\left(261 - \frac{729}{3}\right) = .3333$$

Note. Assuming only a small number of clusters are selected from a large number of clusters in the population, the finite population correction (fpc) is ignored in the computation.